

Application of Expression Profiling to the Developing Lung*

Identification of Putative Regulatory Networks Controlling Matrix Production

Thomas A. Neff Lecture

Thomas J. Mariani, PhD; and
Steven D. Shapiro, MD, FCCP

If we hope to repair damaged lung tissue associated with a variety of acquired and developmental diseases, we must first gain a full appreciation of normal lung development. As an approach, we have utilized Affymetrix (Santa Clara, CA) high-density, oligonucleotide-based microarrays to generate an expression profile of the entire process of rodent lung development, which will be made publicly available. Our initial results were internally consistent and correlated closely with those generated with standard expression techniques such as Northern hybridization. We have verified known expression of genes, found other genes with previously unsuspected expression during lung development, as well as uncovered many expressed sequence tags whose role in lung development awaits further study. Data mining reveals close relationships of expression profiles between specific genes, suggesting novel regulatory relationships. In the future, application of these methods to the study of gene-targeted mice with abnormal lung development should uncover pathways of airway and alveolar development. Ultimately, expression profiling of diseased lungs might allow us to understand why the lung fails to repair, and strategies to influence repair might become apparent.

(*CHEST* 2002; 121:42S–44S)

Abbreviations: cDNA = complementary DNA; cRNA = complementary RNA; ECM = extracellular matrix; EST = expressed sequence tag

Genome-wide expression analysis allows the interrogation of many if not all messenger RNAs expressed in a given cell or tissue. Improvements in technology and widespread use have made this technique affordable for

*From the Departments of Pediatrics, Medicine, Cell Biology and Physiology, and the Program in Lung Development, Washington University School of Medicine and St. Louis Children's Hospital, St. Louis, MO.

This work was supported by the National Heart, Lung, and Blood Institute; Francis Families Foundation; and the Washington University School of Medicine Program in Lung Development. Correspondence to: Stephen D. Shapiro, MD, FCCP, Pulmonary and Critical Care Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis St, Boston MA; e-mail: sshapiro@rics.bwh.harvard.edu

individual laboratories. These types of experiments challenge the hypothesis-driven method that investigators traditionally use to approach biological problems. In contrast, this process-driven approach should lead to a variety of testable hypotheses without suffering from difficulties in predicting candidate genes. This technology has largely been applied as a discriminate tool to determine differences between two conditions. For example, one can determine differences in gene expression in a cell type with and without treatment, or one can compare normal and diseased tissue.¹ Examination of tissues as opposed to cells has additional complexity in accounting for different cell types associated with different samples. Another application of expression profiling is to compare gene expression over time.² Here, the technology can be used as a correlative tool in order to identify genes with similar expression patterns across the time series. Genes sharing similarities in expression often are functionally related.

Publishing databases of large microarray studies in traditional journal format is difficult. Indeed, as more expression databases are constructed, communication of results will become a major issue. Traditional journals struggle with the nonhypothesis-driven approaches and incomplete description of massive amounts of data. In-depth investigation into specific aspects of the data will conform to usual types of publication. In lieu of or in addition to traditional publication, young investigators that immerse themselves in these altruistic tasks must be rewarded in academic circles in other ways. For example, in addition to publications, Web sites (and number of hits or major advances based on the data) should be incorporated into curriculum vitas.

METHODS

Two general formats are widely used. Complementary DNA (cDNA) microarrays contain long nucleic-acid probes immobilized on membranes or glass slides. The distinguishing features of cDNA microarrays are that the expression values are based on the competitive hybridization of two samples being directly compared, and a single hybridization event for each gene/probe. Conversely, oligonucleotide-based microarrays utilize a noncompetitive strategy, where each sample is hybridized independently. This technology is dominated by Affymetrix Inc. (Santa Clara, CA) and their GeneChip arrays. This is the technology that will be the focus of this report.

Technically, total (> 5 µg) or polyA RNA (> 0.2 µg) isolated from each sample is used to generate a biotinylated "target" complementary RNA (cRNA). This is performed in a two-step process, beginning with linear (or amplified) generation of a cDNA library and ending with *in vitro* transcription of this cDNA into cRNA. Following fragmentation of the cRNA, to improve hybridization to the short oligonucleotide "probes" immobilized on the chip, hybridization of each individual sample is performed. Each chip consists of hundreds of thousands of short oligonucleotides, representing between approximately 7,000 to 13,000 genes. Each gene is represented as a "probe set" for which individual expression data are generated, and multiple probe sets

Table 1—A Summary of Clustering Data Highlighting Relationships Between Regulatory Molecules and ECM Molecules*

ECM Component	Known Regulator of Lung Development	Putative Regulator of Lung Development
Elastic fiber (tropoelastin, LTBP, fibulin)	TGF- β , PDGF, LKLF	MFH1
Interstitial collagen (type I, III)	COUP-TF	SOX
Basement membrane collagen (type IV)	FGFR3, FGFR4	NRC-1
Basement membrane, noncollagen (laminin, entactin, fibronectin)	Capsulin/cor1/pod1	

*Note the inclusion of numerous molecules previously implicated in lung development as well as novel pathways; these data suggest potential mechanisms for their actions. TGF = transforming growth factor; PDGF = platelet-derived growth factor; LKLF = lung Kruppel-like factor; COUP-TF = chicken oralbumin upstream promoter transcription factor; FGFR = fibroblast growth factor receptor; MFH = mesenchymal fork-head; SOX = Sry box; NRC = nuclear regulator coactivator; LTBP = latent transforming growth factor-binding protein.

can interrogate the same gene. Each probe set consists of 16 to 20 pairs of oligonucleotides complementary to overlapping segments of an expressed sequence (cloned gene or expressed sequence tag [EST]). One of each pair of oligonucleotides has the correct sequence, while the other contains a single mismatched nucleotide. This strategy is utilized in order to control for random hybridization, with the mismatch oligonucleotide serving as a baseline. For each probe set (gene), evaluation for expression is based on the hybridization characteristics of all 16 to 20 sets of probe pairs (32 to 40 hybridization events).

For comparison purposes, the output must be scaled, usually by one of two strategies: (1) using a small group of "house-keeping" genes that show invariant expression, or (2) using a transcriptome-equivalent strategy, with the assumption that the total sum of all transcripts are similar between samples. Obviously, each strategy has its limitations, but the transcriptome-equivalent strategy is currently more commonly used. After scaling, hard data are generated for each probe set (and each probe, though these data are typically not evaluated individually). Multiple values are generated in an effort to fully describe the expression characteristics of each probe set. No single value gives a complete picture of the data set, yet for intuitive simplicity the relative expression value (average difference value) is most often used. Additional metrics can be used to further describe the data. For instance, the absent/present call may indicate if a given gene was or was not expressed in a sample.

APPROACH

We have applied gene expression profiling to whole-lung tissue throughout lung development. Our intent was to generate a profile of normal mouse development that will serve as a resource of gene expression information and a baseline for future analysis of murine models of abnormal development. Lung development was assessed using Mu11Kchipset subA and subB oligonucleotide microarrays (Affymetrix) as described.³ This high-density oligonucleotide array encompasses > 11,000 cloned genes and ESTs.

Initial analysis was performed on Swiss Webster mice due to their large lung size throughout early lung development. Future applications will include other strains, both to observe similarities and differences between

strains, and to have a firm grasp of lung development in strains used for gene targeting (such as C57BL/6-J). Our approach was to combine multiple (three or more) lungs and perform a single microarray analysis. Tissue was obtained every 2 to 3 days from E12-P14. Adult lung tissue was also obtained. Total cellular RNA was isolated, and 10 μ g was used to generate target cRNA for hybridization to the chip.

The initial strategy of pooling multiple lungs for a single array per time point was based on our experience with the low technical variability of this system as opposed to large biological variability. Data evaluation using a variety of internal controls, such as multiple probe sets for some individual genes (fibronectin, α_1 [I] procollagen) which gave similar expression values, supported the accuracy of our data set. More importantly, application of Northern hybridization to a small number of cloned genes demonstrated a high degree of concordance between the microarray data and traditional expression techniques. Additionally, data mining revealed genes that clustered together most closely in their developmental expression profile are genes that are of the same family or closely related functions. In the future, these experiments will be repeated to test biological and technical variability.

EXTRACELLULAR MATRIX GENE EXPRESSION DURING LUNG DEVELOPMENT

Our initial focus has been on the expression of genes encoding proteins that comprise the extracellular matrix (ECM). ECM formation is an essential component to lung development and repair. The ECM can serve as a structural support for organ morphogenesis, regulate cellular activity with structural cues, and modulate growth factor availability or activity.⁴ Examination of groups of ECM genes shows similar profiles for molecules sharing functional classification.³ Large-scale mathematical clustering^{5,6} of the entire data set also reveals expression profile similarities among genes sharing functional roles. For instance, groups of genes encoding interstitial collagens clustered together, using both hierarchical and agglomerative methods. Basement membrane collagens clustered together with a distinct profile. Included in the collagen nodes were other genes with similar expression patterns including

ESTs of unknown function and known transcription factors (Table 1). Coordinate regulation of transcription factors with ECM proteins leads to hypotheses regarding regulation of ECM gene expression during lung development

CONCLUSIONS

Expression profiling of lung development is a useful tool capable of simultaneously identifying the expression patterns of large numbers of genes and ESTs. These data should serve as a clearinghouse of information for investigators interested in knowing the expression pattern of any specific gene or group of genes in the lung. Even in the context of the dynamic changes in cell populations of whole lung tissue, this technique consistently reported similar expression profiles for genes with previously known functional similarities. Data mining of this massive data set can generate novel, testable hypothesis related to the regulation of lung development. Further experimentation will be essential to validate the specific hypotheses generated by these approaches, as well as the utility of this method to identify regulatory networks essential to the process of lung development.

REFERENCES

- 1 Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286:531–537
- 2 Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998; 9:3273–3297
- 3 Mariani T, Reed J, Shapiro S. Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix. *Am J Respir Cell Mol Biol*. (in press)
- 4 Lukashev M, Werb Z. ECM signalling: orchestrating cell behaviour and misbehaviour. *Trends Cell Biol* 1998; 8:437–441
- 5 Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns *Proc Natl Acad Sci U S A* 1998; 95:14863–14868
- 6 Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* 1999; 96:6745–6750

Transcriptional Profiling of Non-small Cell Lung Cancer Using Oligonucleotide Microarrays*

Gady Cojocaru, BsC; Nir Friedman, PhD; Meir Krupsky, MD; Penina Yaron, MsC; David Simansky, MD, FCCP; Alon Yellin, MD; Gideon Rechavi, MD; Yossef Barash, MsC; Amir Ben-Dor, PhD; Zohar Yakhini, PhD; and Naftali Kaminski, MD

(*CHEST* 2002; 121:44S)

Lung cancer is a common malignancy and a major determinant of overall cancer mortality in developed and developing countries. Despite intensive research, little has changed in the understanding and management of the disease. In order to determine the transcriptional programs that are active in non-small cell lung cancer, we analyzed gene expression patterns using GeneChip U95A microarrays (Affymetrix; Santa Clara, CA) that allow for the analysis of approximately 12,000 genes in 12 non-small cell lung cancer tumor samples, 6 normal histology samples from lung resections for cancer, and pooled normal lung RNA (five individual lungs) obtained commercially. Preliminary analysis revealed that gene expression patterns were highly distinct in tumor and normal tissues. Furthermore, hierarchical clustering clearly distinguished between normal and tumor samples. In order to determine the most informative genes in our data set, we implemented the total-number-of-misclassifications, information-content, and Gaussian-error scores. One evident observation was that informative genes were overabundant in our data set, thus supporting the significance of the results. Among the genes that were most significantly increased in the tumors, we distinguished several categories: genes probably related to cellular infiltrate, such as lymphocyte and macrophage restricted genes; genes clearly related to cancer, such as known oncogenes and cell cycle regulators; and extracellular matrix-related genes possibly representing fibrous tissue. In the genes that were decreased, a symmetrical but opposite trend was observed in cancer-related genes, with known tumor-suppressor genes and inhibitors of cell-cycle progression being decreased. The wealth of statistically significant and biologically meaningful information in our data set supports our contention that transcriptional profiling will lead to new insights into the pathogenesis of lung cancer, thus leading to development of new tools for early detection and treatment of this devastating disease.

*From Functional Genomics (Mr. Cojocaru and Dr. Simansky), Respiratory Medicine (Drs. Krupsky, Kaminski, and Ms. Yaron), Thoracic Surgery (Drs. Yellin and Simansky), and Pediatric Hemato-oncology (Dr. Rechavi), Sheba Medical Center, Tel Hashomer, Israel; Computer Sciences (Dr. Friedman and Mr. Barash), Hebrew University, Jerusalem, Israel; and Agilent Laboratories (Drs. Ben-Dor and Yakhini), Palo Alto, CA. Correspondence to: Naftali Kaminski, MD, Functional Genomics Unit, Molecular Hemato-oncology and Institute of Respiratory Medicine, Hematology Lab Bldg, Room 202, Sheba Medical Center, Tel Hashomer 52621, Israel